Connections across scientific publications based on semantic annotations

Castro, Leyla Jael Garcia ; Berlanga, Rafael ; Rebholz-Schuhmann, Dietrich ; Garcia, Alexander

Abstract: Abstract. The core information from scientific publications is encoded in natural language text and monolithic documents; therefore it is not well integrated with other structured and unstructured data resources. The text format requires additional processing to semantically interlink the publications and to finally reach interoperability of contained data. Data infrastructures such as the Linked Open Data initiative based on the Resource Description Framework support the connectivity of data from scientific publications once the identification of concepts and relations has been achieved, and the content has been interconnected semantically. In this manuscript we produce and analyze the semantic annotations in scientific articles to investigate on the interconnectivity across the articles. In our initial experiment based on articles from PubMed Central we demonstrate the means and the results leading to the interconnectivity using annotations of Medical Subject Headings concepts, Unified Medical Language System terms, and semantic abstractions of relations. We conclude that the different methods would contribute to different types of relatedness between articles that could be later used in recommendation systems based on semantic links across a network of scientific publications.

Originally published at:
Castro, Leyla Jael Garcia; Berlanga, Rafael; Rebholz-Schuhmann, Dietrich; Garcia, Alexander (2013). Connections across scientific publications based on semantic annotations. In: 3rd Workshop on Semantic Publishing (SePublica 2013), 10th Extended Semantic Web Conference, Montpellier, France, 26 May 2013, 51-62.

Connections across scientific publications based on semantic annotations

Leyla Jael Garcia Castro[1], Rafael Berlanga[1], Dietrich Rebholz-Schuhmann[2], Alexander Garcia[3]

[1] Temporal Knowledge Bases Group, Department of Computer Languages and Systems, Universitat Jaume I, Castello de la Plana, Spain
[2] Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3] Institute for Digital Information & Scientific Communication, College of Communication and Information, Florida State University, Tallahassee, Florida, USA

leylajael@gmail.com, berlanga@lsi.uji.es, rebholz@cl.uzh.ch, alexgarciac@gmail.com

Abstract. The core information from scientific publications is encoded in natural language text and monolithic documents; therefore it is not well integrated with other structured and unstructured data resources. The text format requires additional processing to semantically interlink the publications and to finally reach interoperability of contained data. Data infrastructures such as the Linked Open Data initiative based on the Resource Description Framework support the connectivity of data from scientific publications once the identification of concepts and relations has been achieved, and the content has been interconnected semantically. In this manuscript we produce and analyze the semantic annotations in scientific articles to investigate on the interconnectivity across the articles. In our initial experiment based on articles from PubMed Central we demonstrate the means and the results leading to the interconnectivity using annotations of Medical Subject Headings concepts, Unified Medical Language System terms, and semantic abstractions of relations. We conclude that the different methods would contribute to different types of relatedness between articles that could be later used in recommendation systems based on semantic links across a network of scientific publications.

Keywords: Semantic publication, semantic integration and interoperability, life sciences, semantic annotations, concept recognition.

1 Introduction

Scientific publications have traditionally been the primary means by which scholars communicate their work, e.g., new reporting on hypotheses, methods, results, experiments, etc. [1]. New technologies have introduced changes in the handling of scientific publications; however, the knowledge embedded in such documents remains, to a large extent, poorly exploited and interconnected with other data. The reference section relates scientific articles in an explicit way to other scientific documents, i.e., the prior art [2]. Further relatedness results from shared authors and bibliographic metadata. By contrast, all other connectivity based on the knowledge

representation in the content is underexploited, despite the availability of standardized public resources such as the Medical Subject Headings (MeSH) [3], the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [4], and the Unified Medical Language System (UMLS) [5]. These resources would contribute to the construction of knowledge databases facilitating access to semantically normalized information provided by scientific publications.

In this manuscript, we explore on the connections across scientific articles based on their semantic features and annotations. We address the problem of identifying relations between semantic annotations and their relevance for the connectivity between related manuscripts. We examine eleven full-text articles from the open-access subset of PubMed Central and determine which connectivity results from MeSH and UMLS concept annotations. This paper is organized as follows. In Section 2 we introduce our approach while in Section 3 we present the experiment we have carried on, detailing materials, methods, and results. In Section 4 we discuss results and contrast them with related work. Finally in Section 5 we present conclusions and future work.

2 From conceptual features to semantic interconnectivity

We base our approach on the fact that documents do share semantics according to the terminology from the documents. Identifying and annotating terminology has been achieved by different projects, for example the Collaborative Annotation of a Large Biomedical Corpus (CALBC) [6, 7] project. Within CALBC the automatic generation of a large-scale text corpus annotated with biomedical entities, particularly chemical entities, drugs, genes, proteins, diseases, disorders, and species has been studied and the results have been transformed into a triple store [7]. Furthermore, the Nature Publishing Group (NPG) recently released metadata for its publications as Resource Description Framework (RDF) statements; the dataset includes MeSH terms. Finally, the Semantic Enrichment of the Scientific Literature (SESL) [8] project explored the use of semantic web standards and technologies in order to enrich the content of scientific publications: it focused on the integration and interoperability of public and proprietary data resources.

In order to facilitate semantic integration and interoperability for scientific publications, Biotea [9] has built a semantic layer upon the open-access full-text PubMed Central (PMC) articles by transforming the articles into RDF. Biotea also identifies biological entities in the content and abstracts using text-mining and entity-recognition tools, particularly the NCBO Annotator [10] and Whatizit [11, 12]. The identified entities are exposed in RDF as annotations following the model proposed by the Annotation Ontology (AO) [13]. The sets of semantic annotations from the scientific publications facilitate semantic analysis of the unstructured content from the literature.

We augmented the Biotea annotation infrastructure by adding UMLS annotations and by extracting relations involving semantic annotations. In order to identify and semantically categorize these relations, we used several solutions: ReVerb (http://reverb.cs.washington.edu/), a Natural Language Processing (NLP) approach

for relation identification; the Concept Mapping Annotator (CMA) [14]; and a novel semantic-based relation extractor [15]. Both CMA and the relation extractor make use of UMLS, which is one of the most comprehensive knowledge resources in the biomedical domain. Its meta-thesaurus (version 2012AB) covers more than 2.5 million of concepts from over 150 terminological resources, including Medical Subject Headings, NCI Thesaurus, and some others also used for annotations in Biotea. We use the UMLS annotations for the standardization of annotations as well as for the clustering of annotations according to UMLS categories, i.e., the semantic types from the semantic network in UMLS.

In addition and for the future, we propose to include elements of the discourse structure from each manuscript after they have been identified by the SAPIENTA annotator [16]. The relevant Core Scientific Concepts (CoreSC) are labeled as hypothesis, motivation, goal, object, background, method, experiment, model, observation, result and conclusion. Our approach is illustrated in Fig. 1; our main goal is to provide an analytical framework that takes advantage of the semantic features contained in the scientific publications, and focuses on the semantic connections between papers for further information retrieval, recommendation systems and literature-based discovery.

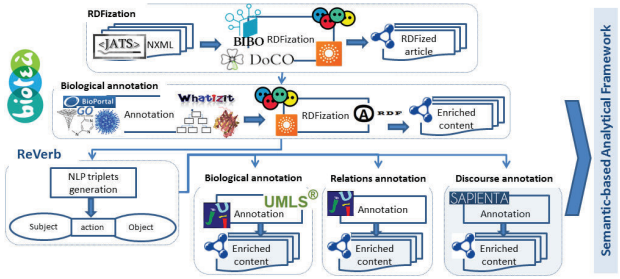

Fig. 1. Our semantic enrichment process. We combine text-mining, entity-recognition, NLP, and semantic techniques in order to provide a semantic layer for scientific publications. The Sapienta components have been shaded, since no results from preliminary experiment will be shown.

3 The augmented Biotea approach

With our analytical framework we have performed an experiment that determines how scientific manuscripts relate to each other based on the co-location of semantic annotations; our analysis relies on concept-based clustering of documents.

3.1 Materials and Methods

From the Biotea SPARQL endpoint (http://biotea.idiginfo.org/query.php), we selected six articles at random from three journals: one from BMC Emergence Medicine, one from Bioinformatics, and four from BMC Biology. All articles satisfy the condition that each one references at least one other manuscript in the endpoint (i.e., ∀x ∃y | x bibo:cites y). In addition to these six articles, we selected five of the referenced articles. Fig. 2 shows the eleven selected articles as well as the SPARQL query which retrieved the initial six documents; Table 1 gives an overview on the selected articles. The process we followed is presented in Fig. 3.

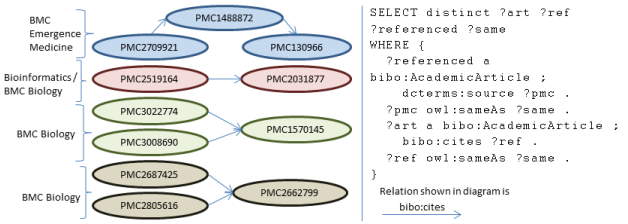

Fig. 2. Selected articles. The SPARQL query included in the figure was used to retrieve the first 100 articles according to the conditions; from them, we selected six referencing at least another one in the subset, and then five of their references. Journals are distinguished by colors.

Table 1. Additional information for selected articles. The five most frequent terms correspond to annotations in Biotea with the highest occurrence in the manuscript.

Articles	Description	Five most frequent terms
PMC 130966	Observational study on patients with suspected acute coronary syndrome (ACS) analyzing characteristics, dispositions, and outcome among patients in order to identify possible improvements in diagnostics.	Patients, ACS, risk, study, symptoms
PMC 1488872	Study on direct hospital costs of chest pain patients in an emergency department (ED)	Patients, cost, ACS, pain, chest
PMC 1570145	Analysis on all genes from sequenced plastid genomes in order to obtain a measure of the overall extent of horizontal gene transfer (HGT) to the plastid	Plastid, genes, HGT, sequence, red
PMC 2031877	Analysis based on the optical projection tomography technique in order to understand how different organ systems and anatomical structures develop throughout the life of	Zebrafish, development, OPT, model, data

	the zebrafish	
PMC 2519164	Review on advances on molecular and cellular microscopic images in bioinformatics, including applications, techniques, tools and resources	Image, analysis, patterns, techniques, data
PMC 2662799	Study baed on full-length sequences of transcripts for *Buchnera aphidicola* and *Acyrthosiphon pisum*, and detailed structural and phylogenetic analyses in order to assess the possibility of lateral gene transfer	Genes, *buchnera*, *ldcA* (gene), *rlpA* (gene), bacteria
PMC 2687425	Commentary on the evolutionary importance of the transfer of genes between host and symbiont	Genes, transfers, genome, host, lateral
PMC 2709921	Evaluation on utility and costs of acute nuclear myocardial perfusion imaging (MPI) in an ED for patients with suspected ACS.	MPI, patients, ACS, cost, study
PMC 2805616	Analysis on integration of non-retroviral ribonucleic acid (RNA) virus genes on fungal hosts, and function of those genes. It uses sequencing across host-virus gene boundaries and phylogenetic analyses of fungal hosts and totivirids	Genes, totivirus, viral, RdRp (RNA polymerase), RNA
PMC 3008690	Summary of two studies related to patterns, processes, and consequences of HGT	Gene, plant, HGT, conversion, gene conversion
PMC 3022774	Analysis on the extent and evolutionary fate of HGT in the parasitic genus *Cuscuta* and a small clade of *Plantago* species aiming to understand details on the mechanics for plant-to-plant HGT	Mitochondrial, transfer, DNA, *plantago*, *atp1* (gene)

In the first step, we retrieved from Biotea the RDF data for the semantic annotations, and selected only those annotations referring to MeSH concepts. We also collected the MeSH terms assigned to the manuscripts in PubMed. In this way we were able to analyze how articles related to each other based on the co-occurrence of MeSH concepts for both datasets, Biotea and PubMed; these first steps correspond to processes 1 and 2 in Fig. 3. From Biotea, we selected the sections and paragraphs, as illustrated by the third step in Fig. 3. We applied ReVerb to the paragraph sections in order to identify sentences that comply with the syntactic form (subject, predicate, object), step 4. As we are interested in the concepts contained in the sentences, we did not discard any sentence at this point of the analysis. For the sentences (see step 5) we applied another annotation tool, called CMA [12]. Similar to the NCBO Annotator and Whatizit, CMA identifies biological entities; furthermore, CMA associates the identified entities with Concept Unique Identifiers (CUIs) from UMLS Meta-thesaurus. Both NCBO Annotator and Whatizit use a dictionary-based text-mining technique while CMA –similar to MetaMap [17]– applies concept classification techniques to stretches of text. For CMA a user may select a threshold to specify the

minimum level of confidence. In our case, we used a low setting to induce high recall. The annotations from CMA contributed in a second analysis towards the relatedness measurements for scientific articles based on the co-occurrence of UMLS terms.

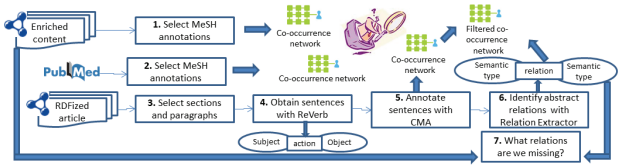

Fig. 3. Our method at a glance.

CMA identifies the subject and object in a sentence, and it is then possible to use [15] in order to identify and categorize the relation between two CUIs (step 6). The semantic relation extractor proposed by [15] extracts the relations between any pair of CUIs in the annotated sentences, resolves synonyms, and produces semantic clusters where the relations are grouped according to synonymous mentions of concepts; in short, it generates abstractions of relations. Such abstract relation form templates where both subject and object make reference to UMLS semantic types; as a result, the template can be applied to any pair of CUIs belonging to the identified semantic types. These identified abstract relations are the basis to the third relatedness analysis.

As we are only receiving relations for those sentences where both subject and object are annotated; we analyzed the annotations provided by Biotea, i.e., the annotations from the NCBO Annotator and Whatizit, for sentences that were not processed by [15]. This way, we could determine the number of relations that went missing (step 7 in Fig. 3). We split the sentences with zero or one recognized CUIs into three subsets: in the first both the subject and the object have been annotated in Biotea, in the second, either subject or object has been annotated, and in the third, all remaining sentences are kept. Below we show the formalization of these subsets in the Formula 1. The first set would tell us exactly how many possible relations we are missing, the second set shows us relations that can be retrieved from ontologies –even when we have identified only one concept in the sentence, other sentences can have ontologically related terms, and the third set contains sentences without enough information for relation extraction.

$$\begin{aligned} &S = \{x \mid x=<\text{subject,action,object}>\} = A \cup B \cup C,\ \emptyset = A \cap B \cap C \\ &A=\{x \mid x=<\text{subject,action,object}> \wedge \text{isAnnotated(subject)} \wedge \text{isAnnotated(object)},\ x \in S\} \\ &B=\{x \mid x=<\text{subject,action,object}> \wedge (\text{isAnnotated(subject)} \veebar \text{isAnnotated(object)}),\ x \in S\} \\ &C=\{x \mid x=<\text{subject,action,object}> \wedge \neg\text{isAnnotated(subject)} \wedge \neg\text{isAnnotated(object)}),\ x \in S\} \end{aligned} \quad (1)$$

3.2 Results

Eleven manuscripts have been annotated with Biotea, CMA, and our semantic relation extractor [15]. In total, the data set comprised 340 paragraphs from 171 sections. We

identified a total of 2088 sentences with ReVerb from which only 1232 had CUIs for both subject and object. From these sentences, a total of 261 abstract relations were extracted. Table 2 gives a summary of our data set.

Table 2. Our working set.

Articles	Sections	Paragraphs	ReVerb Sentences	Analyzed Sentences	Abstract Relations
PMC130966	18	34	150	81	15
PMC1488872	21	30	161	88	27
PMC1570145	26	45	330	172	30
PMC2031877	13	25	164	83	6
PMC2519164	23	51	236	119	3
PMC2662799	21	34	301	177	72
PMC2687425	2	8	73	45	19
PMC2709921	14	31	163	107	25
PMC2805616	10	24	209	122	28
PMC3008690	4	9	56	26	2
PMC3022774	19	49	364	213	34
TOTALS	171	340	2207	1233	261

For the first analysis we examined the connections across the articles based on MeSH concepts. Table 3 presents a summary of the MeSH annotations retrieved from Biotea and PubMed; it also includes the UMLS annotations retrieved with CMA as well as the relations with highest confidence. Annotations from Biotea were retrieved with a SPARQL query while annotations from PubMed were manually gathered. As we were interested in the relatedness between articles, we analyzed the shared annotations that are defined as any concept being referenced as an annotation both in publication *A* and *B*. From the shared annotations we moved to shared concepts, i.e., biological entities associated with a unique entry in a controlled vocabulary.

Table 3. MeSH and UMLS concepts and relation examples in our working set.

Articles	Biotea Mesh	PubMed Mesh	CMA UMLS	Relations with highest confidence
PMC130966	85	Not found	301	Discharged, improved, suitable
PMC1488872	73	Not found	291	Discharged, admitted, defined
PMC1570145	92	8	626	Flanked, matched, extracted
PMC2031877	49	8	302	Examine, prevented, represents
PMC2519164	103	9	466	Begun, attracted, fused
PMC2662799	93	20	799	Encoded, enter, transferred
PMC2687425	28	6	145	Express, functional, reveal
PMC2709921	91	17	330	Discharged, participated, identify
PMC2805616	75	21	361	Encode, integrated, function
PMC3008690	29	8	142	Propose, leads
PMC3022774	79	17	550	Adjacent, converted, enabled

We found 783 shared concepts in Biotea and 33 in PubMed. As the number of shared concepts from Biotea was much higher than the number from PubMed, we selected only concepts with a weight greater than 1.0. The weights varied from 0.04 to 10.41; a total of 67 shared concepts were above the chosen threshold. The weight for shared concepts is defined in Formula 2; as a same concept can be annotated with multiple terms, for instance both "gene" and "genes" could be annotated with the concept MeSH-D005796, we summed up the occurrences by concept rather than term.

$$\text{weight(shared concept C)} = \frac{\left(\frac{\#of\ occurrences\ of\ C}{\#\ of\ sections}\ in\ article\ A + \frac{\#of\ occurrences\ of\ C}{\#\ of\ sections}\ in\ article\ B\right)}{2} \quad (2)$$

Fig. 4 depicts the connections based on MeSH concepts for Biotea and PubMed. The articles corresponding to BMC Emergence Medicine journal were clustered separately from the rest. This is not clearly visible in the graph that corresponds to annotations from PubMed. We did not find MeSH annotations for the articles PMC130966 and PMC1488772; thus, these two articles together with PMC2709921 are isolated in this graph. In both cases, Biotea and PubMed, PMC2687425 is the most connected article; it has relations to six articles. However, it is not connected to PMC2031877 in Biotea, and to PMC2519164 in PubMed. PMC3022774 is also connected to six articles in Biotea but only to five in PubMed; it is not connected to PMC2031877 in Biotea, and to PMC2031877 and PMC2519164 in PubMed. Surprisingly, although PMC2519164 cites PMC2031877 they are not connected by MeSH concepts.

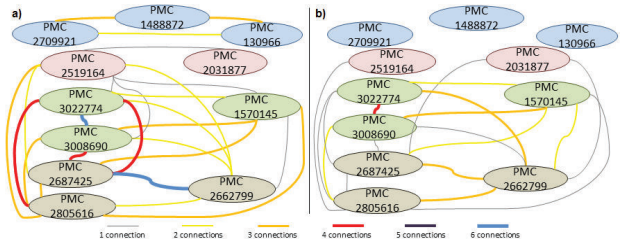

Fig. 4. a) Connections between articles based on MeSH concepts from Biotea. b) Connections between articles based on PubMed MeSH concepts.

Similar to the analysis performed for MeSH terms, we also examined the annotations obtained with CMA. As the threshold was low, we got a large number of terms and concepts; therefore, we also applied the weight formula for CMA annotations. The weights varied from 0.002 to 0.24; we selected the weight 0.1 as cutting point. We found a total of 2343 annotations with CMA covering 2429 different concepts. However, the number could be higher had we annotated the entire

article and not just the sentences identified with ReVerb. Fig. 5 shows the connections according to UMLS concepts from CMA and the extracted relations, i.e., without and with a relation-based filter applied. Similar to the connections from MeSH terms, PMC130966, PMC1488772, and PMC2709921 shaped an independent cluster. However in this occasion PMC1488772 is also connected to PMC2031877, the connections come from the concept "model". The rest of the articles are grouped in a second cluster; there PMC2687425 is connected to all the other articles but PMC2031877, same as it happens in Biotea. Same as it happened in PubMed, PMC3022774 is not connected to either PMC2031877 or PMC2519164. Different as it happened from MeSH connections, this time PMC2687425 is connected to PMC2031877 (indeed the former cites the latter). Fig. 5b shows the same relations but with a relation-based filter applied. From the extracted abstract relations, we chose one "discharge" that takes subjects from the UMLS semantic type T001[LIVB] and objects from T061[PROC]. For PMC1570145 and PMC3008690 no annotation with a semantic type T001 was found.

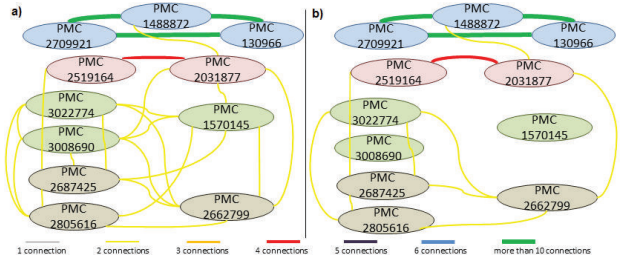

Fig. 5. a) Connections between articles based on UMLS concepts from CMA. b) Connections between articles based on UMLS concepts from CMA with a relation-based filter applied.

Finally, we analyzed the sentences where CUIs were identified only for either the subject or the object, or none of them at all, i.e., for 855 of 2088 retrieved sentences by ReVerb. As illustrated in Formula 1, we partitioned these 975 sentences in three sets: (i) the set A –CUIs for subject and object, with 339 sentences, (ii) the set B –CUIs for either the subject or the object, with 487 sentences; and (iii) the set C –no CUI identified, with 29 sentences. From the relations extractor, we originally got 261 relations from 1233 sentences; corresponding to the 21%. Assuming (i) a linear relation between the number of the sentences with CUIs for both (subject, object) and the abstract relations retrieved, and (ii) only new abstract relations would be retrieved from these 339 new sentences. Then we would be missing about 71 new relations. In relation to the total this is still the 20%. Even when 20% seems to be low, it is important to note that abstract relations actually cover more than one relation as a UMLS semantic type can be applied to multiple CUIs.

4 Discussion

Articles naturally relate to each other via citations; articles sharing citations are considered similar to some extent [2]. Text-based approaches such as term frequency-inverse document frequency and latent semantic analysis have also been used to measure similarity across documents [18]. In a similar vein, cluster-based approaches have also been explored. For instance, Lewis [19] groups related articles by using a keyword-based method followed by a sentence-alignment algorithm that ranks and orders the initial results. Similarly, McSyBi [20] clusters articles according to a set of topics; the information for the creation of topics is gathered from titles and abstracts. Different from Lewis, McSyBi enables the use of MeSH terms or UMLS semantic types in order to modify the clusters so that users can analyze the data from different perspectives.

Unlike these approaches, we are working with a semantically annotated dataset in contrast to plain text articles. Similar to McSyBi, we use MeSH and UMLS concepts in order to calculate relatedness between articles, and explore on the opportunities from semantic annotations of full-text documents. We are reporting connections that later could lead to a semantic-based similarity model for scientific publications. Connections based on MeSH concepts were similar in both cases, Biotea and PubMed. This indicates that it is indeed possible to define a semantic-based approach to measure relatedness across articles. For our working set we only found connections not inferred from PubMed MeSH annotations for those articles without reported MeSH concepts. We cannot conclude yet whether semantic relatedness would be more or less accurate than the relatedness implicit in the related articles suggested in PubMed. However, the similarities in both graphs are a good starting point to extend our relatedness approach to more specific annotations that could introduce a difference; for instance, proteins, genes, diseases, drugs, among others. UMLS connections graph also exhibits similarities with those coming from MeSH terms; therefore, it seems feasible to use other vocabularies, and combining them, in order to find a similarity measure between articles.

Different ways of narrowing the initial connections are possible. For our sample, the relation-based filter applied to the connections did not represent a significant difference. However, it could be improved by using also ontological relations. Even though only connections to two articles were excluded, we still consider that this filtering is a possibility worth exploring further. Rather than using the filtering only for exclusion, it could be also incorporated in the relatedness formula. Although we have explored only the connections across articles, there are other possibilities that can be built on top of semantic dataset for scientific publications as the one provided by Biotea and the extension we propose in this paper. In the biomedical domain, several authors have reported different methods aiming to find hidden relations from semantic annotations. For instance, from MeSH terms it is possible to identify patterns that can be used to find candidates for new associations between drugs and diseases [21]. Similarly, recognizing Gene Ontology terms co-occurring with human gene can be used to discover possible Gene Ontology annotations for those genes [22]. Also, the identification of shared annotations across genes can contribute to identify possible relationships between those genes [23, 24].

5 Conclusions and future work

We have explored how articles connect to each other from a semantic perspective. We have evaluated different concept annotation solutions on full text documents to determine to which extend relatedness can be inferred from such annotations. Such relatedness should facilitate to automatically and semantically integrate literature into an infrastructure of interlinked data elements. Although this semantic-based relatedness project is still in its initial stage, the results from our preliminary experiment are promising. We have found that connections across articles from annotations automatically identified with entity recognition tools, e.g., Whatizit, NCBO Annotator, and CMA, are similar to those connections exhibit based on the PubMed MeSH terms. Having sematic annotations for other vocabularies opens new and interesting possibilities. For instance, it becomes possible to analyze the connections from different perspectives i.e., different vocabularies as well as combinations of them. Additionally, we have also shown the use of relation-based filters in order to narrow the found connections from the co-occurrence of concepts. In our case, we used abstract relations extracted from those sentences where both subject and object were identified by CMA; however, it is also possible to use the relations coming from the ontologies. Different analysis can be performed on the sentences with only one or no biological entities identified; not necessarily about relatedness but also about hidden relations in the plain text.

As part of our future work we have considered to (i) improve the input for ReVerb so we can get more accurate sentences, (ii) use CMA to annotate the entire corpus as it was done in Biotea with the NCBO Annotator and Whatizit, (iii) use relations from the ontologies used to annotated the corpus, (iv) improve our initial weight formula, (v) integrate discourse-based annotations from SAPIENTA, and (vi) formalize a semantic-based method to measure relatedness across scientific publications. The discourse elements provided by SAPIENTA will be used to filter the relations depending on whether or not the participating concepts are related to a particular set of scientific concepts; such set would be define by users.

References

1. Swan, A.: Overview of scholarly communication. In: Jacobs, N. (ed.): Open Access: Key Strategic, Technical and Economic Aspects. Chandos (2006)
2. Hummon, N.P., Dereian, P.: Connectivity in a citation network: The development of DNA theory. Social Networks 11 (1989) 39-63
3. Rogers, F.: Medical subject headings. Bulletin of the Medical Library Association 51 (1963) 114-116
4. Cornet, R., de Keizer, N.: Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 8 Suppl 1 (2008) S2
5. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32 (2004) D267-D270
6. Rebholz-Schuhmann, D., Yepes, A., Van Mulligen, E., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC silver standard corpus. Bioinformatics and computational biology 8 (2010) 163-179

7. Croset, S., Grabmüller, C., Li, C., Kavaliauskas, S., Rebholz-Schuhmann, D.: The CALBC RDF Triple Store: retrieval over large literature content. International Workshop on Semantic Web Applications and Tools for the Life Sciences, Berin, Germany (2010)
8. Harrow, I., Filsell, W., Woollard, P., Dix, I., Braxenthaler, M., Gedye, R., Hoole, D., Kidd, R., Wilson, J., Rebholz-Schuhmann, D.: Towards Virtual Knowledge Broker services for semantic integration of life science literature and data sources. Drug Discovery Today in press (2012)
9. Garcia Castro, L.J., McLaughlin, C., Garcia, A.: Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data. Biomedical semantics 4 Suppl 1 (2013) S5
10. Jonquet, C., Shah, N.H., Youn, C.H., Callendar, C., Storey, M.-A., Musen, M.A.: NCBO Annotator: Semantic Annotation of Biomedical Data. International Semantic Web Conference, Poster and Demo session (2009)
11. Rebholz-Schuhmann, D., Arregui, M., Gaudan, M., Kirsch, H., Jimeno, A.: Text processing through Web Services: Calling Whatizit. Bioinformatics 24 (2007) 296-298
12. Kirsch, H., Rebholz-Schuhmann, D.: Distributed modules for text annotation and IE applied to the biomedical domain. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland (2004) 50-53
13. Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics 2 (2011) S4
14. Berlanga, R., Nebot, V., Jimenez-Ruiz, E.: Semantic annotation of biomedical texts through concept retrieval. Procesamiento de Lenguaje Natural 45 (2010) 247-250
15. Nebot, V., Berlanga, R.: Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. Knowledge and information Systems (2012) 1-25
16. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28 (2012) 991-1000
17. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17 (2010) 229-236
18. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., Börner, K.: Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE 6 e18029
19. Lewis, J., Ossowski, S., Hicks, J., Errami, M., Garner, H.R.: Text similarity: an alternative way to search MEDLINE. Bioinformatics 22 (2006) 2298-2304
20. Yamamoto, Y., Takagi, T.: Biomedical knowledge navigation by literature clustering. Journal of Biomedical Informatics 40 (2007) 114-130
21. Srinivasan, P., Libbus, B., Sehgal, A.K.: Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases. In: Hirschman, L., Pustejovsky, J. (eds.): Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL, Boston, Massachusetts, USA (2004) 33-40
22. Good, B., Su, A.I.: Mining Gene Ontology Annotations From Hyperlinks in the Gene Wiki. Translational Bioinformatics Conference, Washington, D.C. (2011)
23. Saha, B., Hoch, A., Khuller, S., Raschid, L., Zhang, X.-N.: Dense subgraphs with restrictions and applications to gene annotation graphs. 14th Annual international conference on Research in Computational Molecular Biology, Vol. 6044. Springer-Verlag, Lisbon, Portugal (2010) 456-472
24. Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.-N.: Link prediction for annotation graphs using graph summarization. International Conference on the Semantic Web, Vol. 7031. Springer-Verlag, Bonn, Germany (2011) 714-729